Differentiability on $\left(\mathbb{R}^{n\times m}, \mathbb{R}^{d\times n\times m}\right)$

Orestis Ioannou

December 2024

Contents

I Part I: Foundations and Basic Theory **5**

1 The Metric on $\mathbb{R}^{n\times m}$ **5**

2 Convergence and Completeness **6**

3 Partial Derivatives **7**

4 The Inner Product on $\mathbb{R}^{n\times m}$ **8**

5 Differentiability of Scalar-Valued Functions **8**

6 Continuity of Scalar-Valued Functions **10**

7 Algebraic Properties of Differentiability **10**

8 Matrix-Valued Functions and Tensor Notation **12**

9 Component-Wise Characterization of Matrix Differentiability **12**

10 Summation Operator **13**

11 Hadamard-Type Product **14**

12 Continuity of Matrix-Valued Functions **14**

13 Norm Inequality for Contraction Product **15**

14 Boundedness of Difference Quotients **16**

15 Chain Rule: Scalar-Valued Composition **16**

16 Chain Rule: Matrix-Valued Composition **17**

17 Scalar Composition: Differentiable Outer Function **18**

18 Corollary: Elementwise Composition **18**

19 Application 1: Right Multiplication by Constant Matrix **19**

20 Application 2: Left Multiplication by Constant Matrix **20**

21 Application 3: Broadcasting (Outer Product) **20**

22 Application: Deep Neural Networks **20**

II Differentiability on $\mathbb{R}^{d\times n\times m}$ 23

23 Partial Derivatives and the Jacobian on $\mathbb{R}^{d\times n\times m}$ **23**

24 Inner Product on $\mathbb{R}^{d\times n\times m}$ **23**

25 Differentiability of Scalar-Valued Functions on $\mathbb{R}^{d\times n\times m}$ **24**

26 Continuity on $\mathbb{R}^{d\times n\times m}$ **25**

27 Algebra of Differentiable Functions on $\mathbb{R}^{d\times n\times m}$ **25**

28 Tensor Spaces and Operations on $\left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$ **26**

29 The Summation Operator and Hadamard Product on 3-Tensors **27**

30 Boundedness Under Differentiability on $\mathbb{R}^{d\times n\times m}$ **28**

31 The Chain Rule on $\mathbb{R}^{d\times n\times m}$ (Scalar Case) **29**

32 The Scalar Chain Rule on $\mathbb{R}^{d\times n\times m}$ **30**

33 Element-wise Composition on $\mathbb{R}^{d\times n\times m}$ **31**

34 Slicing, Submatrix Extraction, and Convolution Operations **32**

35 Application: Convolution with Fixed Filter **32**

36 Application: Convolution with Fixed Input **33**

37 Bias Addition **34**

38 Application to Convolutional Neural Networks **34**
38.1 Gradient with respect to D 34
38.2 Gradient with respect to ω 35
38.3 Gradient with respect to b 35

Part I

Part I: Foundations and Basic Theory

1 The Metric on $\mathbb{R}^{n\times m}$

We begin by equipping $\mathbb{R}^{n\times m}$ with the Frobenius norm, following standard constructions in analysis [Rud76, HJ12].

Definition 1.1. Let $A, B \in \mathbb{R}^{n\times m}$. We define the norm

$$\|A - B\| = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \left(A_{(i,j)} - B_{(i,j)} \right)^2 \right)^{1/2}$$

Proposition 1.1. Let $A, B, C \in \mathbb{R}^{n\times m}$. Then:

1. $\|A - B\| \geq 0$
2. $\|A - B\| = \|B - A\|$
3. $\|A - B\| = 0 \iff A = B$
4. $\|A - B\| \leq \|A - C\| + \|C - B\|$ (triangle inequality)

Proof. Items 1, 2, and 3 are immediate.

For item 4:

$$\begin{aligned} \|A - B\| &= \left(\sum_i \sum_j (A_{(i,j)} - B_{(i,j)})^2 \right)^{1/2} \\ &= \left(\sum_i \sum_j (A_{(i,j)} - C_{(i,j)} + C_{(i,j)} - B_{(i,j)})^2 \right)^{1/2} \end{aligned}$$

Expanding the square:

$$\begin{aligned} = \Bigg(\sum_i \sum_j \big[&(A_{(i,j)} - C_{(i,j)})^2 + (C_{(i,j)} - B_{(i,j)})^2 \\ &+ 2(A_{(i,j)} - C_{(i,j)})(C_{(i,j)} - B_{(i,j)}) \big] \Bigg)^{1/2} \end{aligned}$$

By the Cauchy–Schwarz inequality [Cau21, Sch85] applied to the cross term:

$$\leq \left(\sum_i \sum_j (A_{(i,j)} - C_{(i,j)})^2 + \sum_i \sum_j (C_{(i,j)} - B_{(i,j)})^2 \right.$$
$$\left. +2 \left(\sum_i \sum_j (A_{(i,j)} - C_{(i,j)})^2 \right)^{1/2} \left(\sum_i \sum_j (C_{(i,j)} - B_{(i,j)})^2 \right)^{1/2} \right)^{1/2}$$

Recognizing this as a perfect square:

$$= \left(\left[\left(\sum_i \sum_j (A_{(i,j)} - C_{(i,j)})^2 \right)^{1/2} + \left(\sum_i \sum_j (C_{(i,j)} - B_{(i,j)})^2 \right)^{1/2} \right]^2 \right)^{1/2}$$
$$= \|A - C\| + \|C - B\|$$

□

Corollary 1.1. The space $(\mathbb{R}^{n\times m}, \|\cdot\|)$ is a metric space.

2 Convergence and Completeness

Definition 2.1. Let $A : \mathbb{N} \to \mathbb{R}^{n\times m}$ and $\hat{A} \in \mathbb{R}^{n\times m}$. We say that

$$\lim_{n\to\infty} A_n = \hat{A} \iff \forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : n \geq n_0 \implies \|A_n - \hat{A}\| < \varepsilon$$

Proposition 2.1. Let $A : \mathbb{N} \to \mathbb{R}^{n\times m}$ and $\hat{A} \in \mathbb{R}^{n\times m}$. The following are equivalent:

1. $\lim_{n\to\infty} A_n = \hat{A}$
2. For all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$: $\lim_{n\to\infty} A_n(i,j) = \hat{A}(i,j)$

Proof. ($1 \implies 2$): Let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Then:

$$|A_n(i,j) - \hat{A}(i,j)| \leq \left(\sum_i \sum_j (A_n(i,j) - \hat{A}(i,j))^2 \right)^{1/2} = \|A_n - \hat{A}\| \to 0$$

$(2 \implies 1)$: If all component sequences converge, then:

$$\lim_{k\to\infty}\left(\sum_i\sum_j (A_k(i,j)-\hat{A}(i,j))^2\right)^{1/2}=0$$

so $A_n \to \hat{A}$. □

Corollary 2.1. $(\mathbb{R}^{n\times m}, \|\cdot\|)$ is complete.

Proof. Let $A:\mathbb{N}\to\mathbb{R}^{n\times m}$ be a Cauchy sequence. For each $i\in\{1,\dots,n\}$ and $j\in\{1,\dots,m\}$, define the scalar sequence $a_{(i,j)}:\mathbb{N}\to\mathbb{R}$ by $a_{(i,j)}(n)=A_n(i,j)$.

Since A is Cauchy, each component sequence $a_{(i,j)}$ is Cauchy in $\mathbb{R}$. Because $(\mathbb{R},|\cdot|)$ is complete, there exists $\hat{a}_{(i,j)}\in\mathbb{R}$ such that $\lim_{n\to\infty} a_{(i,j)}(n)=\hat{a}_{(i,j)}$.

Define $\hat{A}\in\mathbb{R}^{n\times m}$ by $\hat{A}(i,j)=\hat{a}_{(i,j)}$. Then $\lim_{n\to\infty}A_n=\hat{A}$ by the previous proposition. □

3 Partial Derivatives

Definition 3.1. Let $\Omega\subseteq\mathbb{R}^{n\times m}$ be open, $f:\Omega\to\mathbb{R}$, and $A\in\Omega$. We say that f is partially differentiable at A in the direction $\hat{e}_{(i,j)}$ if

$$\exists\lambda\in\mathbb{R}:\forall\varepsilon>0,\exists\delta>0:0<|h|<\delta \implies \left|\frac{f(A+h\hat{e}_{(i,j)})-f(A)}{h}-\lambda\right|<\varepsilon$$

Equivalently, for any sequence $h:\mathbb{N}\to\mathbb{R}\setminus\{0\}$ with $A+h_n\hat{e}_{(i,j)}\in\Omega$ for all n and $h_n\to 0$, we have

$$\frac{f(A+h_n\hat{e}_{(i,j)})-f(A)}{h_n}\to\lambda$$

We write $\frac{\partial f}{\partial\hat{e}_{(i,j)}}(A)=\lambda$.

Observation 3.1. Let $\Omega\subseteq\mathbb{R}^{n\times m}$ be open, $f:\Omega\to\mathbb{R}$, and $A\in\Omega$. The following are equivalent:

1. $\frac{\partial f}{\partial\hat{e}_{(i,j)}}(A)$ exists.

2. For any sequence $h : \mathbb{N} \to \mathbb{R} \setminus \{0\}$ with $A + h_n \hat{e}_{(i,j)} \in \Omega$ and $h_n \to 0$, we have
$$\lim_{n\to\infty} \frac{f(A + h_n \hat{e}_{(i,j)}) - f(A)}{h_n} = \frac{\partial f}{\partial \hat{e}_{(i,j)}}(A)$$

Definition 3.2. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f : \Omega \to \mathbb{R}$, and $A \in \Omega$. We say that f has all partial derivatives at A if for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, the partial derivative $\frac{\partial f}{\partial \hat{e}_{(i,j)}}(A)$ exists.

Define the Jacobian matrix $J_f(A) \in \mathbb{R}^{n\times m}$ by
$$J_f(A) = \begin{bmatrix} \frac{\partial f}{\partial \hat{e}_{(1,1)}}(A) & \cdots & \frac{\partial f}{\partial \hat{e}_{(1,m)}}(A) \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial \hat{e}_{(n,1)}}(A) & \cdots & \frac{\partial f}{\partial \hat{e}_{(n,m)}}(A) \end{bmatrix}$$

4 The Inner Product on $\mathbb{R}^{n\times m}$

Definition 4.1. Let $A, B \in \mathbb{R}^{n\times m}$. We define the inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^{n\times m} \times \mathbb{R}^{n\times m} \to \mathbb{R}$ by
$$\langle A, B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{(i,j)} B_{(i,j)}$$

5 Differentiability of Scalar-Valued Functions

Definition 5.1. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f : \Omega \to \mathbb{R}$, and $A \in \Omega$. We say that f is differentiable at A if
$$\exists M \in \mathbb{R}^{n\times m} : \forall \varepsilon > 0, \exists \delta > 0 : 0 < \|B - A\| < \delta \implies \frac{|f(B) - f(A) - \langle B - A, M \rangle|}{\|B - A\|} < \varepsilon$$

Equivalently, for any sequence $B : \mathbb{N} \to \Omega \setminus \{A\}$ with $B_n \to A$, we have
$$\frac{|f(B_n) - f(A) - \langle B_n - A, M \rangle|}{\|B_n - A\|} \to 0$$

Observation 5.1. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f : \Omega \to \mathbb{R}$, and $A \in \Omega$. The following are equivalent:

1. f is differentiable at A.
2. There exists $M \in \mathbb{R}^{n\times m}$ such that for any sequence $B : \mathbb{N} \to \Omega \setminus \{A\}$ with $B_n \to A$, we have
$$\frac{|f(B_n) - f(A) - \langle B_n - A, M\rangle|}{\|B_n - A\|} \to 0$$

Proposition 5.1. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f : \Omega \to \mathbb{R}$, and $A \in \Omega$ with f differentiable at A. Then for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$,
$$M_{(i,j)} = \frac{\partial f}{\partial \hat{e}_{(i,j)}}(A)$$
Thus $M = J_f(A)$.

Proof. Let $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Let $h : \mathbb{N} \to \mathbb{R} \setminus \{0\}$ be any sequence with $A + h_n\hat{e}_{(i,j)} \in \Omega$ and $h_n \to 0$. Set $B_n = A + h_n\hat{e}_{(i,j)}$. Then $B_n \to A$, so
$$\frac{|f(B_n) - f(A) - \langle B_n - A, M\rangle|}{\|B_n - A\|} \to 0$$
Now:
$$\frac{|f(B_n) - f(A) - \langle B_n - A, M\rangle|}{\|B_n - A\|} = \frac{|f(A + h_n\hat{e}_{(i,j)}) - f(A) - h_nM_{(i,j)}|}{|h_n|} \to 0$$
This implies:
$$\left|\frac{f(A + h_n\hat{e}_{(i,j)}) - f(A)}{h_n} - M_{(i,j)}\right| \to 0$$
Therefore, $\frac{\partial f}{\partial \hat{e}_{(i,j)}}(A) = M_{(i,j)}$. □

Proposition 5.2. For all $A, B \in \mathbb{R}^{n\times m}$,
$$|\langle A, B\rangle| \le \|A\|\|B\|$$

Proof.
$$\begin{aligned}
|\langle A, B\rangle| &= \left|\sum_i \sum_j A_{(i,j)}B_{(i,j)}\right| \\
&\le \left(\sum_i \sum_j A^2_{(i,j)}\right)^{1/2} \left(\sum_i \sum_j B^2_{(i,j)}\right)^{1/2} \qquad \text{(Cauchy–Schwarz)} \\
&= \|A\|\|B\|
\end{aligned}$$

□

6 Continuity of Scalar-Valued Functions

Definition 6.1. Let $X \subseteq \mathbb{R}^{n\times m}$, $f : X \to \mathbb{R}$, and $A \in X$. We say that f is continuous at A if

$$\forall \varepsilon > 0, \exists \delta > 0 : B \in X, \|B - A\| < \delta \implies |f(A) - f(B)| < \varepsilon$$

Remark 6.1. f is continuous at A if and only if for any sequence $B : \mathbb{N} \to X$ with $B_n \to A$, we have $\lim f(B_n) = f(A)$.

Proposition 6.1. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f : \Omega \to \mathbb{R}$, and $A \in \Omega$. If f is differentiable at A, then f is continuous at A.

Proof. Let $B : \mathbb{N} \to \Omega$ be any sequence with $B_n \to A$. Define $\varphi : \mathbb{N} \to \mathbb{R}$ by

$$\varphi_n = \begin{cases} \frac{|f(B_n)-f(A)|}{\|B_n-A\|} & \text{if } B_n \neq A \\ 0 & \text{otherwise} \end{cases}$$

Then $|f(B_n) - f(A)| = \varphi_n \|B_n - A\|$.
Claim: φ is bounded.
Proof of claim: By differentiability of f at A,

$$\begin{aligned} |f(B_n) - f(A)| &\leq \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \cdot \|B_n - A\| + |\langle B_n - A, J_f(A)\rangle| \\ &\leq \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} + \|B_n - A\| \cdot \frac{\|J_f(A)\|}{\|B_n - A\|} \\ &= \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} + \|J_f(A)\| \end{aligned}$$

Since $\lim_n \frac{|f(B_n)-f(A)-\langle B_n-A,J_f(A)\rangle|}{\|B_n-A\|} = 0$, the sequence φ is bounded.
Therefore, $|f(B_n) - f(A)| = \varphi_n \|B_n - A\| \to 0$, so f is continuous at A. □

7 Algebraic Properties of Differentiability

Proposition 7.1. Let $\Omega \subseteq \mathbb{R}^{n\times m}$ be open, $f, g : \Omega \to \mathbb{R}$, $\lambda \in \mathbb{R}$, and $A \in \Omega$ with f and g differentiable at A. Then:

1. $(f+g):\Omega\to\mathbb{R}$ is differentiable at A, and $J_{(f+g)}(A)=J_f(A)+J_g(A)$.
2. $(\lambda f):\Omega\to\mathbb{R}$ is differentiable at A, and $J_{(\lambda f)}(A)=\lambda J_f(A)$.
3. $(f\cdot g):\Omega\to\mathbb{R}$ is differentiable at A, and $J_{(fg)}(A)=f(A)J_g(A)+g(A)J_f(A)$.

Proof. Let $B:\mathbb{N}\to\Omega\setminus\{A\}$ be any sequence with $B_n\to A$.

(1)

$$\frac{|(f+g)(B_n)-(f+g)(A)-\langle B_n-A,J_f(A)+J_g(A)\rangle|}{\|B_n-A\|}$$
$$\leq\frac{|f(B_n)-f(A)-\langle B_n-A,J_f(A)\rangle|}{\|B_n-A\|}+\frac{|g(B_n)-g(A)-\langle B_n-A,J_g(A)\rangle|}{\|B_n-A\|}\to 0$$

(2)

$$\frac{|(\lambda f)(B_n)-(\lambda f)(A)-\langle B_n-A,\lambda J_f(A)\rangle|}{\|B_n-A\|}=|\lambda|\frac{|f(B_n)-f(A)-\langle B_n-A,J_f(A)\rangle|}{\|B_n-A\|}\to 0$$

(3) We have:

$$\frac{|(fg)(B_n)-(fg)(A)-\langle B_n-A,f(A)J_g(A)+g(A)J_f(A)\rangle|}{\|B_n-A\|}$$
$$=\frac{|f(B_n)(g(B_n)-g(A))+g(A)(f(B_n)-f(A))-\langle B_n-A,f(A)J_g(A)\rangle-\langle B_n-A,g(A)J_f(A)\rangle|}{\|B_n-A\|}$$
$$\leq\frac{|f(B_n)(g(B_n)-g(A))-\langle B_n-A,f(A)J_g(A)\rangle|}{\|B_n-A\|}+|g(A)|\frac{|f(B_n)-f(A)-\langle B_n-A,J_f(A)\rangle|}{\|B_n-A\|}$$

The second term vanishes by differentiability of f. For the first term, denote it as a_n and decompose:

$$a_n=\frac{|f(B_n)(g(B_n)-g(A))-\langle B_n-A,f(B_n)J_g(A)\rangle+\langle B_n-A,(f(B_n)-f(A))J_g(A)\rangle|}{\|B_n-A\|}$$
$$\leq\frac{|f(B_n)(g(B_n)-g(A))-\langle B_n-A,f(B_n)J_g(A)\rangle|}{\|B_n-A\|}+\frac{|\langle B_n-A,(f(B_n)-f(A))J_g(A)\rangle|}{\|B_n-A\|}$$

The first part vanishes by factoring out $f(B_n)$ (which is continuous, hence bounded). For the second part:

$$\frac{|\langle B_n-A,(f(B_n)-f(A))J_g(A)\rangle|}{\|B_n-A\|}\leq\frac{\|B_n-A\||f(B_n)-f(A)|\|J_g(A)\|}{\|B_n-A\|}=|f(B_n)-f(A)|\|J_g(A)\|\to$$

Thus $a_n\to 0$, completing the proof. □

8 Matrix-Valued Functions and Tensor Notation

Definition 8.1. An element $A \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$ is an $n \times m$ matrix whose entries are $\tilde{n} \times \tilde{m}$ matrices:

$$A = \begin{bmatrix} A_{(1,1)} & \cdots & A_{(1,m)} \\ \vdots & \ddots & \vdots \\ A_{(n,1)} & \cdots & A_{(n,m)} \end{bmatrix} \quad \text{where each } A_{(i,j)} \in \mathbb{R}^{\tilde{n}\times\tilde{m}}$$

Definition 8.2. Let $B \in \mathbb{R}^{\tilde{n}\times\tilde{m}}$ and $A \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$. We define the contraction product $B \odot A \in \mathbb{R}^{n\times m}$ by

$$(B \odot A)_{(i,j)} = \langle B, A_{(i,j)} \rangle$$

Definition 8.3. Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ be open and $f : \Omega \to \mathbb{R}^{n\times m}$, where

$$f(x) = \begin{bmatrix} f_{(1,1)}(x) & \cdots & f_{(1,m)}(x) \\ \vdots & \ddots & \vdots \\ f_{(n,1)}(x) & \cdots & f_{(n,m)}(x) \end{bmatrix}$$

and each $f_{(i,j)} : \Omega \to \mathbb{R}$.

Let $A \in \Omega$. We say that f is differentiable at A if there exists $M_f \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$ such that

$$\forall \varepsilon > 0, \exists \delta > 0 : B \in \Omega, \|B-A\| < \delta \implies \frac{\|f(B) - f(A) - (B-A) \odot M_f\|}{\|B-A\|} < \varepsilon$$

Equivalently, for any sequence $B : \mathbb{N} \to \Omega \setminus \{A\}$ with $B_n \to A$,

$$\frac{\|f(B_n) - f(A) - (B_n - A) \odot M_f\|}{\|B_n - A\|} \to 0$$

9 Component-Wise Characterization of Matrix Differentiability

Proposition 9.1. Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ be open, $f : \Omega \to \mathbb{R}^{n\times m}$, and $A \in \Omega$. Then:

1. If f is differentiable at A, then for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, the component function $f_{(i,j)}$ is differentiable at A, and

$$M_f = \begin{bmatrix} J_{f_{(1,1)}}(A) & \cdots & J_{f_{(1,m)}}(A) \\ \vdots & \ddots & \vdots \\ J_{f_{(n,1)}}(A) & \cdots & J_{f_{(n,m)}}(A) \end{bmatrix}$$

2. If all component functions $f_{(i,j)}$ are differentiable at A, then f is differentiable at A.

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$ be any sequence with $B_n \to A$.

(1) Let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Then:

$$\frac{|f_{(i,j)}(B_n) - f_{(i,j)}(A) - \langle B_n - A, M_f(i,j)\rangle|}{\|B_n - A\|} \leq \frac{\|f(B_n) - f(A) - (B_n - A) \odot M_f\|}{\|B_n - A\|} \to 0$$

Therefore, $f_{(i,j)}$ is differentiable at A and $J_{f_{(i,j)}}(A) = M_f(i,j)$.

(2) Define

$$M_f = \begin{bmatrix} J_{f_{(1,1)}}(A) & \cdots & J_{f_{(1,m)}}(A) \\ \vdots & \ddots & \vdots \\ J_{f_{(n,1)}}(A) & \cdots & J_{f_{(n,m)}}(A) \end{bmatrix}$$

Then:

$$\frac{\|f(B_n) - f(A) - (B_n - A) \odot M_f\|}{\|B_n - A\|}$$

$$= \left(\sum_i \sum_j \frac{(f_{(i,j)}(B_n) - f_{(i,j)}(A) - \langle B_n - A, J_{f_{(i,j)}}(A)\rangle)^2}{\|B_n - A\|^2} \right)^{1/2} \to 0$$

Thus f is differentiable at A. □

10 Summation Operator

Definition 10.1. Let $A \in (\mathbb{R}^{\tilde{n} \times \tilde{m}})^{n \times m}$. Define the summation operator $\Sigma : (\mathbb{R}^{\tilde{n} \times \tilde{m}})^{n \times m} \to \mathbb{R}^{\tilde{n} \times \tilde{m}}$ by

$$\Sigma(A) = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{(i,j)}$$

Proposition 10.1. Let $A, B \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$ and $\lambda \in \mathbb{R}$. Then:

1. $\Sigma(A+B) = \Sigma(A) + \Sigma(B)$
2. $\Sigma(\lambda A) = \lambda\Sigma(A)$

Proof. These follow directly from properties of summation. □

11 Hadamard-Type Product

Definition 11.1. Let $A \in \mathbb{R}^{n\times m}$ and $B \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$. We define the Hadamard-type product $A * B \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$ by

$$(A * B)_{(i,j)} = A_{(i,j)}B_{(i,j)}$$

where the multiplication on the right is scalar multiplication of a scalar by a matrix.

Explicitly:

$$A * B = \begin{bmatrix} A_{(1,1)}B_{(1,1)} & \cdots & A_{(1,m)}B_{(1,m)} \\ \vdots & \ddots & \vdots \\ A_{(n,1)}B_{(n,1)} & \cdots & A_{(n,m)}B_{(n,m)} \end{bmatrix}$$

12 Continuity of Matrix-Valued Functions

Definition 12.1. Let $X \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$, $f : X \to \mathbb{R}^{n\times m}$, and $A \in X$. We say that f is continuous at A if

$$\forall \varepsilon > 0, \exists \delta > 0 : B \in X, \|B - A\| < \delta \implies \|f(B) - f(A)\| < \varepsilon$$

Remark 12.1. f is continuous at A if and only if for any sequence $B : \mathbb{N} \to X$ with $B_n \to A$, we have $f(B_n) \to f(A)$.

Proposition 12.1. Let $X \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$, $f : X \to \mathbb{R}^{n\times m}$, and $A \in X$. Then f is continuous at A if and only if for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, the component function $f_{(i,j)}$ is continuous at A.

Proof. Let $B : \mathbb{N} \to X$ be any sequence with $B_n \to A$.

($\Rightarrow$): Let $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$. Then:

$$|f_{(i,j)}(B_n) - f_{(i,j)}(A)| \leq \|f(B_n) - f(A)\| \to 0$$

($\Leftarrow$): If all components are continuous, then:

$$\|f(B_n) - f(A)\| = \left(\sum_i \sum_j (f_{(i,j)}(B_n) - f_{(i,j)}(A))^2 \right)^{1/2} \to 0$$

□

Corollary 12.1. Let $\Omega \subseteq \mathbb{R}^{\tilde{n} \times \tilde{m}}$ be open, $f : \Omega \to \mathbb{R}^{n \times m}$, and $A \in \Omega$. If f is differentiable at A, then f is continuous at A.

Proof. Since f is differentiable at A, each component $f_{(i,j)}$ is differentiable at A, hence continuous at A. Therefore, f is continuous at A by the previous proposition. □

13 Norm Inequality for Contraction Product

Proposition 13.1. Let $A \in \mathbb{R}^{\tilde{n} \times \tilde{m}}$ and $B \in (\mathbb{R}^{\tilde{n} \times \tilde{m}})^{n \times m}$. Then

$$\|A \odot B\| \leq \|A\| \|B\|$$

Proof. We have:

$$\begin{aligned} \|A \odot B\| &= \left(\sum_i \sum_j (A \odot B)_{(i,j)}^2 \right)^{1/2} \\ &= \left(\sum_i \sum_j \langle A, B_{(i,j)} \rangle^2 \right)^{1/2} \\ &= \left(\sum_i \sum_j \left(\sum_k \sum_l A_{(k,l)} B_{(i,j)}(k,l) \right)^2 \right)^{1/2} \end{aligned}$$

By Cauchy–Schwarz applied to each term:

$$\begin{aligned}
&\leq \left(\sum_i \sum_j \left(\sum_k \sum_l A_{(k,l)}^2\right)\left(\sum_k \sum_l B_{(i,j)}(k,l)^2\right)\right)^{1/2} \\
&= \left(\sum_k \sum_l A_{(k,l)}^2\right)^{1/2} \left(\sum_i \sum_j \sum_k \sum_l B_{(i,j)}(k,l)^2\right)^{1/2} \\
&= \|A\|\|B\|
\end{aligned}$$

□

14 Boundedness of Difference Quotients

Proposition 14.1. Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ be open, $f : \Omega \to \mathbb{R}^{n\times m}$, and $A \in \Omega$ with f differentiable at A. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$ be any sequence with $B_n \to A$. Then the sequence

$$a_n = \frac{\|f(B_n) - f(A)\|}{\|B_n - A\|}$$

is bounded.

Proof. By differentiability and the contraction product norm inequality:

$$\begin{aligned}
a_n &\leq \frac{\|f(B_n) - f(A) - (B_n - A) \odot J_f(A)\|}{\|B_n - A\|} + \frac{\|(B_n - A) \odot J_f(A)\|}{\|B_n - A\|} \\
&\leq \frac{\|f(B_n) - f(A) - (B_n - A) \odot J_f(A)\|}{\|B_n - A\|} + \|J_f(A)\|
\end{aligned}$$

The first term vanishes as $n \to \infty$ by differentiability, so a_n is bounded. □

15 Chain Rule: Scalar-Valued Composition

Theorem 15.1 (Chain Rule)**.** Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ and $G \subseteq \mathbb{R}^{n\times m}$ be open. Let $f : \Omega \to G$ and $g : G \to \mathbb{R}$. Suppose $A \in \Omega$, f is differentiable at A, and g is differentiable at $f(A)$. Then the composition $g \circ f : \Omega \to \mathbb{R}$ is differentiable at A and

$$J_{g\circ f}(A) = \Sigma(J_g(f(A)) * J_f(A))$$

Here, $J_g(f(A)) \in \mathbb{R}^{n\times m}$ with entries $\frac{\partial g}{\partial \hat{e}_{(i,j)}}(f(A))$, and $J_f(A) \in (\mathbb{R}^{\tilde{n}\times\tilde{m}})^{n\times m}$ with entries $J_{f_{(i,j)}}(A)$.

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$ be any sequence with $B_n \to A$. We decompose the difference:

$$\frac{|(g\circ f)(B_n) - (g\circ f)(A) - \langle B_n - A, \Sigma(J_g(f(A)) * J_f(A))\rangle|}{\|B_n - A\|}$$
$$= \frac{|g(f(B_n)) - g(f(A)) - \langle B_n - A, \Sigma(J_g(f(A)) * J_f(A))\rangle|}{\|B_n - A\|}$$

Write $g(f(B_n)) - g(f(A)) = \sum_i \sum_j \frac{\partial g}{\partial \hat{e}_{(i,j)}}(f(A))(f_{(i,j)}(B_n) - f_{(i,j)}(A)) +$ higher order terms.

Let $\alpha_n = \frac{\|f(B_n)-f(A)-(B_n-A)\odot J_f(A)\|}{\|B_n-A\|}$ and $\beta_n = \frac{\|g(f(B_n))-g(f(A))-(f(B_n)-f(A))\odot J_g(f(A))\|}{\|f(B_n)-f(A)\|}$.

Then $\alpha_n \to 0$ by differentiability of f, and $\beta_n \to 0$ by differentiability of g. Moreover, the sequence $\frac{\|f(B_n)-f(A)\|}{\|B_n-A\|}$ is bounded by the previous proposition.

A careful computation shows that the composed difference quotient is controlled by terms involving α_n and β_n, each multiplied by bounded quantities. Therefore, the composition is differentiable with the stated Jacobian. □

16 Chain Rule: Matrix-Valued Composition

Theorem 16.1 (Chain Rule for Matrix-Valued Functions)**.** Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ and $G \subseteq \mathbb{R}^{\hat{n}\times\hat{m}}$ be open. Let $f : \Omega \to G$ and $g : G \to \mathbb{R}^{n\times m}$. Suppose $A \in \Omega$, f is differentiable at A, and g is differentiable at $f(A)$. Then the composition $g \circ f : \Omega \to \mathbb{R}^{n\times m}$ is differentiable at A and

$$J_{(g\circ f)}(A) = \begin{bmatrix} \Sigma(J_{g_{(1,1)}}(f(A)) * J_f(A)) & \cdots & \Sigma(J_{g_{(1,m)}}(f(A)) * J_f(A)) \\ \vdots & \ddots & \vdots \\ \Sigma(J_{g_{(n,1)}}(f(A)) * J_f(A)) & \cdots & \Sigma(J_{g_{(n,m)}}(f(A)) * J_f(A)) \end{bmatrix}$$

Proof. The proof proceeds similarly to the scalar case. Each component $(g_{(i,j)} \circ f)(B_n) - (g_{(i,j)} \circ f)(A)$ is analyzed using the scalar chain rule. The bound on the difference quotient for f (from the previous proposition) ensures that the composed error converges to zero. Collecting components yields the stated Jacobian. □

17 Scalar Composition: Differentiable Outer Function

Proposition 17.1. Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ be open and $f : \Omega \to \mathbb{R}$ with $A \in \Omega$. Suppose f is differentiable at A. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable. Then the composition $g \circ f : \Omega \to \mathbb{R}$ is differentiable at A and

$$J_{g\circ f}(A) = g'(f(A))J_f(A)$$

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$ be a sequence with $B_n \to A$. Decompose:

$$\begin{aligned}&\frac{|g(f(B_n)) - g(f(A)) - g'(f(A))(f(B_n) - f(A))|}{\|B_n - A\|}\\ &= \frac{|g(f(B_n)) - g(f(A)) - g'(f(A))(f(B_n) - f(A))|}{\|f(B_n) - f(A)\|} \cdot \frac{\|f(B_n) - f(A)\|}{\|B_n - A\|}\end{aligned}$$

The first factor vanishes by differentiability of g at $f(A)$. The second factor is bounded by a previous proposition. Thus:

$$\begin{aligned}&\frac{|(g \circ f)(B_n) - (g \circ f)(A) - \langle B_n - A, g'(f(A))J_f(A)\rangle|}{\|B_n - A\|}\\ &\leq \frac{|g(f(B_n)) - g(f(A)) - g'(f(A))(f(B_n) - f(A))|}{\|f(B_n) - f(A)\|} \cdot \frac{\|f(B_n) - f(A)\|}{\|B_n - A\|}\\ &\quad + g'(f(A))\frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \to 0\end{aligned}$$

□

18 Corollary: Elementwise Composition

Corollary 18.1. Let $\Omega \subseteq \mathbb{R}^{\tilde{n}\times\tilde{m}}$ be open, $f : \Omega \to \mathbb{R}^{n\times m}$, and $A \in \Omega$ with f differentiable at A. Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable. Define $g\langle f\rangle : \Omega \to \mathbb{R}^{n\times m}$ by

$$g\langle f\rangle_{(i,j)} = g \circ f_{(i,j)}$$

Then $g\langle f\rangle$ is differentiable at A and

$$J_{g\langle f\rangle}(A) = g'[f(A)] * J_f(A)$$

where $g'[f(A)] \in \mathbb{R}^{n\times m}$ with entries $g'[f(A)]_{(i,j)} = g'(f_{(i,j)}(A))$.

Proof. Apply the scalar composition result to each component $g \circ f_{(i,j)}$:

$$J_{g \circ f_{(i,j)}}(A) = g'(f_{(i,j)}(A)) J_{f_{(i,j)}}(A)$$

Thus:

$$J_{g\langle f \rangle}(A)_{(i,j)} = g'(f_{(i,j)}(A)) J_{f_{(i,j)}}(A) = [g'[f(A)] * J_f(A)]_{(i,j)}$$

□

19 Application 1: Right Multiplication by Constant Matrix

Let $D \in \mathbb{R}^{n_2 \times n_3}$. Define $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_3}$ by $f(x) = x \times D$ (matrix multiplication).

For $x \in \mathbb{R}^{n_1 \times n_2}$,

$$f_{(i,j)}(x) = \sum_{k=1}^{n_2} x_{(i,k)} D_{(k,j)}$$

Thus:

$$\frac{\partial f_{(i,j)}}{\partial \hat{e}_{(\hat{i},\hat{j})}}(x) = \begin{cases} D_{(\hat{j},j)} & \text{if } i = \hat{i} \\ 0 & \text{if } i \neq \hat{i} \end{cases}$$

The Jacobian is:

$$J_f(x) = \begin{bmatrix} \hat{e}_{(1,1)} \otimes D^T & \cdots & \hat{e}_{(1,n_3)} \otimes D^T \\ \vdots & \ddots & \vdots \\ \hat{e}_{(n_1,1)} \otimes D^T & \cdots & \hat{e}_{(n_1,n_3)} \otimes D^T \end{bmatrix}$$

where $\hat{e}_{(i,j)}$ is the standard basis matrix in $\mathbb{R}^{n_1 \times n_3}$ with a 1 in position (i, j) and zeros elsewhere.

For $\Lambda \in \mathbb{R}^{n_1 \times n_3}$:

$$\begin{aligned} \Sigma(\Lambda * J_f(x)) &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_3} \Lambda_{(i,j)} (\hat{e}_{(i,j)} \otimes D^T) \\ &= \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_3} \Lambda_{(i,j)} \hat{e}_{(i,j)} \right) \times D^T \\ &= \Lambda \times D^T \end{aligned}$$

20 Application 2: Left Multiplication by Constant Matrix

Let $W \in \mathbb{R}^{n_1 \times n_2}$. Define $f : \mathbb{R}^{n_2 \times n_3} \to \mathbb{R}^{n_1 \times n_3}$ by $f(x) = W \times x$.

For $x \in \mathbb{R}^{n_2 \times n_3}$,

$$f_{(i,j)}(x) = \sum_{k=1}^{n_2} W_{(i,k)} x_{(k,j)}$$

Thus:

$$\frac{\partial f_{(i,j)}}{\partial \hat{e}_{(\hat{i},\hat{j})}}(x) = \begin{cases} W_{(i,\hat{i})} & \text{if } j = \hat{j} \\ 0 & \text{if } j \neq \hat{j} \end{cases}$$

The Jacobian is:

$$J_f(x) = \begin{bmatrix} W^T \times \hat{e}_{(1,1)} & \cdots & W^T \times \hat{e}_{(1,n_3)} \\ \vdots & \ddots & \vdots \\ W^T \times \hat{e}_{(n_2,1)} & \cdots & W^T \times \hat{e}_{(n_2,n_3)} \end{bmatrix}$$

For $\Lambda \in \mathbb{R}^{n_1 \times n_3}$:

$$\Sigma(\Lambda * J_f(x)) = W^T \times \Lambda$$

21 Application 3: Broadcasting (Outer Product)

Define $f : \mathbb{R}^{n_1 \times 1} \to \mathbb{R}^{n_1 \times n_3}$ by $f(x) = x \times \mathbf{1}^{n_3}$ (repeat column n_3 times).

From Application 2 with $W = I$ (the identity) and $D = \mathbf{1}^{n_3}$ (row vector of ones), we have:

For $\Lambda \in \mathbb{R}^{n_1 \times n_3}$:

$$\Sigma(\Lambda * J_f(x)) = \Lambda \times (\mathbf{1}^{n_3})^T = \Lambda \times \mathbf{1}^{1 \times n_3}$$

which sums all columns of Λ.

22 Application: Deep Neural Networks

We apply the preceding theory to feedforward neural networks [RHW86, GBC16]. Consider a neural network layer with:

- Weight matrix $W \in \mathbb{R}^{n_1 \times n_2}$ (known)
- Data matrix $D \in \mathbb{R}^{n_2 \times n_3}$ (known)
- Bias vector $b \in \mathbb{R}^{n_1 \times 1}$ (known)
- Activation function $\sigma : \mathbb{R} \to \mathbb{R}$ (differentiable)
- Output $Y = \sigma(W \times D + b \times \mathbf{1}^{n_3})$ where σ is applied elementwise
- Loss function $E : \mathbb{R}^{n_1 \times n_3} \to \mathbb{R}$ with known gradient $J_E(Y) \in \mathbb{R}^{n_1 \times n_3}$

Computation of gradient with respect to weights W:
Define:

- $f_1 : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_3}$ by $f_1(x) = x \times D + b \times \mathbf{1}^{n_3}$
- $f_2 = \sigma\langle f_1 \rangle : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_3}$
- $f_W = E \circ f_2 : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$

Then by the chain rule:

$$J_{f_W}(W) = \Sigma(J_E(f_2(W)) * J_{f_2}(W))$$

By the elementwise composition corollary:

$$J_{f_2}(W) = \sigma'[Y] * J_{f_1}(W)$$

where $\sigma'[Y] \in \mathbb{R}^{n_1 \times n_3}$ has entries $\sigma'(Y_{(i,j)})$. For the sigmoid activation, $\sigma'(Y) = Y(1 - Y)$.
Thus:

$$J_{f_W}(W) = \Sigma(J_E(Y) * [\sigma'[Y] * J_{f_1}(W)])$$

From Application 1, $\Sigma(\Lambda * J_{f_1}(W)) = \Lambda \times D^T$. Therefore:

$$\begin{aligned} J_{f_W}(W) &= \Sigma([J_E(Y) \odot \sigma'[Y]] * J_{f_1}(W)) \\ &= [J_E(Y) \odot \sigma'[Y]] \times D^T \quad \boxed{\text{dimension } n_1 \times n_2} \end{aligned}$$

Computation of gradient with respect to data D:
Define:

- $G_1 : \mathbb{R}^{n_2 \times n_3} \to \mathbb{R}^{n_1 \times n_3}$ by $G_1(x) = W \times x + b \times \mathbf{1}^{n_3}$

- $G_2 = \sigma\langle G_1 \rangle$
- $G_D = E \circ G_2$

Then:

$$J_{G_D}(D) = \Sigma(J_E(Y) * J_{G_2}(D)) = \Sigma(J_E(Y) * [\sigma'[Y] * J_{G_1}(D)])$$

From Application 2, $\Sigma(\Lambda * J_{G_1}(D)) = W^T \times \Lambda$. Therefore:

$$J_{G_D}(D) = W^T \times [J_E(Y) \odot \sigma'[Y]] \quad \boxed{\text{dimension } n_2 \times n_3}$$

Computation of gradient with respect to bias b:
Define:

- $h_1 : \mathbb{R}^{n_1 \times 1} \to \mathbb{R}^{n_1 \times n_3}$ by $h_1(x) = W \times D + x \times \mathbf{1}^{n_3}$
- $h_2 = \sigma\langle h_1 \rangle$
- $h_b = E \circ h_2$

Then:

$$J_{h_b}(b) = \Sigma(J_E(Y) * J_{h_2}(b)) = \Sigma(J_E(Y) * [\sigma'[Y] * J_{h_1}(b)])$$

From Application 3, $\Sigma(\Lambda * J_{h_1}(b)) = \Lambda \times \mathbf{1}^{n_3}$, which sums the columns. Therefore:

$$J_{h_b}(b) = [J_E(Y) \odot \sigma'[Y]] \times \mathbf{1}^{n_3} \quad \boxed{\text{dimension } n_1 \times 1}$$

The loss function:
The standard squared error loss is $E_A : \mathbb{R}^{n \times m} \to \mathbb{R}$ defined by

$$E_A(Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{(i,j)} - \hat{Y}_{(i,j)})^2$$

Its gradient is:

$$\frac{\partial E_A}{\partial \hat{e}_{(i,j)}}(Y) = 2(Y_{(i,j)} - \hat{Y}_{(i,j)})$$

Thus:

$$J_{E_A}(Y) = 2(Y - \hat{Y}) \quad \square$$

Part II
Differentiability on $\mathbb{R}^{d\times n\times m}$

23 Partial Derivatives and the Jacobian on $\mathbb{R}^{d\times n\times m}$

Definition 23.1 (Partial Derivative)**.** Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f:\Omega\to\mathbb{R}$, $A\in\Omega$. We say that f is partially differentiable at A in the direction $\hat{e}_{(i,j,k)}$ if:

$$\exists\lambda\in\mathbb{R}:\forall\varepsilon>0:\exists\delta>0:0<|h|<\delta,\ A+h\hat{e}_{(i,j,k)}\in\Omega\implies\left|\frac{f(A+h\hat{e}_{(i,j,k)})-f(A)}{h}-\lambda\right|<\varepsilon$$

Equivalently: $h:\mathbb{N}\to\mathbb{R}\setminus\{0\}$, $\forall n\in\mathbb{N}:A+h_n\hat{e}_{(i,j,k)}\in\Omega$, $h_n\to 0$ implies

$$\frac{f(A+h_n\hat{e}_{(i,j,k)})-f(A)}{h_n}\to\lambda$$

When f is partially differentiable at A in the direction $\hat{e}_{(i,j,k)}$, we write $\lambda=\dfrac{\partial f}{\partial\hat{e}_{(i,j,k)}}(A)$.

Definition 23.2. Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f:\Omega\to\mathbb{R}$, $A\in\Omega$, such that

$$\forall i\in\{1,\dots,d\},\ \forall j\in\{1,\dots,n\},\ \forall k\in\{1,\dots,m\}:\quad\frac{\partial f}{\partial\hat{e}_{(i,j,k)}}(A)\text{ exists.}$$

We set $J_f(A)\in\mathbb{R}^{d\times n\times m}$:

$$J_f(A)_{(i,j,k)}=\frac{\partial f}{\partial\hat{e}_{(i,j,k)}}(A).$$

24 Inner Product on $\mathbb{R}^{d\times n\times m}$

Definition 24.1 (Inner Product)**.** Let $A,B\in\mathbb{R}^{d\times n\times m}$. We define $\langle\cdot,\cdot\rangle:\mathbb{R}^{d\times n\times m}\times\mathbb{R}^{d\times n\times m}\to\mathbb{R}$ by

$$\langle A,B\rangle=\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}A_{(i,j,k)}B_{(i,j,k)}.$$

25 Differentiability of Scalar-Valued Functions on $\mathbb{R}^{d\times n\times m}$

Definition 25.1 (Differentiability)**.** Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f:\Omega\to\mathbb{R}$, $A\in\Omega$. We say that f is differentiable at A if and only if

$$\exists M\in\mathbb{R}^{d\times n\times m}:\forall\varepsilon>0:\exists\delta>0:0<\|B-A\|<\delta \implies \frac{|f(B)-f(A)-\langle B-A,M\rangle|}{\|B-A\|}<\varepsilon.$$

Equivalently: $\exists M\in\mathbb{R}^{d\times n\times m}$ such that for all $B:\mathbb{N}\to\Omega\setminus\{A\}$ with $B_n\to A$:

$$\frac{|f(B_n)-f(A)-\langle B_n-A,M\rangle|}{\|B_n-A\|}\to 0.$$

Proposition 25.1. Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f:\Omega\to\mathbb{R}$, $A\in\Omega$, f differentiable at A. Then $M=J_f(A)$.

Proof. Let $i\in\{1,\dots,d\}$, $j\in\{1,\dots,n\}$, $k\in\{1,\dots,m\}$. Let $h:\mathbb{N}\to\mathbb{R}\setminus\{0\}$: $\forall n\in\mathbb{N}: A+h_n\hat{e}_{(i,j,k)}\in\Omega$, $h_n\to 0$. Set $B:\mathbb{N}\to\Omega\setminus\{A\}$ with $B_n=A+h_n\hat{e}_{(i,j,k)}$. Then $B_n\to A$. Hence

$$\frac{|f(B_n)-f(A)-\langle B_n-A,M\rangle|}{\|B_n-A\|}\to 0$$

which gives

$$\frac{|f(A+h_n\hat{e}_{(i,j,k)})-f(A)-h_nM_{(i,j,k)}|}{|h_n|}\to 0$$

and therefore

$$\left|\frac{f(A+h_n\hat{e}_{(i,j,k)})-f(A)}{h_n}-M_{(i,j,k)}\right|\to 0.$$

□

Proposition 25.2 (Cauchy–Schwarz [Cau21, Sch85])**.** $A,B\in\mathbb{R}^{d\times n\times m}\implies |\langle A,B\rangle|\le\|A\|\|B\|$.

Proof.

$$\begin{aligned}|\langle A,B\rangle| &= \left|\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}A_{(i,j,k)}B_{(i,j,k)}\right|\\ &\le\left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}A_{(i,j,k)}^2\right)^{1/2}\left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}B_{(i,j,k)}^2\right)^{1/2}=\|A\|\|B\|.\end{aligned}$$ □

26 Continuity on $\mathbb{R}^{d\times n\times m}$

Definition 26.1 (Continuity)**.** Let $X \subset \mathbb{R}^{d\times n\times m}$, $f : X \to \mathbb{R}$, $A \in X$. We say f is continuous at A if and only if

$$\forall \varepsilon > 0 : \exists \delta > 0 : B \in X, \|B - A\| < \delta \implies |f(A) - f(B)| < \varepsilon.$$

Equivalently: $B : \mathbb{N} \to X$, $B_n \to A \implies f(B_n) \to f(A)$.

Proposition 26.1. Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f : \Omega \to \mathbb{R}$, $A \in \Omega$, f differentiable at A. Then f is continuous at A.

Proof. Let $B : \mathbb{N} \to \Omega$: $B_n \to A$. Set $\varphi : \mathbb{N} \to \mathbb{R}$ with

$$\varphi_n = \begin{cases} \dfrac{|f(B_n) - f(A)|}{\|B_n - A\|} & B_n \neq A \\ 0 & \text{otherwise.} \end{cases}$$

Then: $|f(B_n) - f(A)| = \varphi_n \|B_n - A\|$.

Claim: φ is bounded.

Indeed:

$$\begin{aligned} |f(B_n) - f(A)| &\leq \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \cdot \|B_n - A\| + \frac{|\langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \cdot \|B_n - A\| \\ &\leq \delta_n \|B_n - A\| + \|J_f(A)\| \|B_n - A\| \end{aligned}$$

where $\delta_n \to 0$. Hence φ is bounded. Therefore $|f(B_n) - f(A)| \to 0$, so f is continuous. □

27 Algebra of Differentiable Functions on $\mathbb{R}^{d\times n\times m}$

Proposition 27.1. Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f, g : \Omega \to \mathbb{R}$, $\lambda \in \mathbb{R}$, $A \in \Omega$, f, g differentiable at A. Then:

1. $(f + g) : \Omega \to \mathbb{R}$ is differentiable at A and $J_{(f+g)}(A) = J_f(A) + J_g(A)$.
2. $(\lambda f) : \Omega \to \mathbb{R}$ is differentiable at A and $J_{(\lambda f)}(A) = \lambda J_f(A)$.
3. $(f \cdot g) : \Omega \to \mathbb{R}$ is differentiable at A and $J_{(fg)}(A) = f(A)J_g(A) + g(A)J_f(A)$.

28 Tensor Spaces and Operations on $\left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$

Definition 28.1. Let $A \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$, i.e., $\forall i \in \{1,\dots,d\},\ \forall j \in \{1,\dots,n\},\ \forall k \in \{1,\dots,m\} : A_{(i,j,k)} \in \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$.

Definition 28.2 ($\odot$ operation)**.** Let $B \in \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $A \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$. Then $B \odot A \in \mathbb{R}^{d\times n\times m}$ with

$$(B \odot A)_{(i,j,k)} = \langle B, A_{(i,j,k)}\rangle.$$

Definition 28.3 (Differentiability of tensor-valued functions)**.** Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : \Omega \to \mathbb{R}^{d\times n\times m}$, $A \in \Omega$.

$$\forall i \in \{1,\dots,d\},\ \forall j \in \{1,\dots,n\},\ \forall k \in \{1,\dots,m\} : f_{(i,j,k)} : \Omega \to \mathbb{R}.$$

We say f is differentiable at A if and only if $\exists M_f \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$:

$$\forall \varepsilon > 0 : \exists \delta > 0 : B \in \Omega, \|B-A\| < \delta \implies \frac{\|f(B) - f(A) - (B-A)\odot M_f\|}{\|B-A\|} < \varepsilon.$$

Equivalently: $\exists M_f$ such that for all $B : \mathbb{N} \to \Omega \setminus \{A\}$ with $B_n \to A$:

$$\lim_{n\to\infty} \frac{\|f(B_n) - f(A) - (B_n - A)\odot M_f\|}{\|B_n - A\|} = 0.$$

Proposition 28.1. Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : \Omega \to \mathbb{R}^{d\times n\times m}$, $A \in \Omega$. Then:

1. If f is differentiable at A, then $\forall i \in \{1,\dots,d\}$, $\forall j \in \{1,\dots,n\}$, $\forall k \in \{1,\dots,m\}$: $f_{(i,j,k)}$ is differentiable at A and $M_f(i,j,k) = J_{f_{(i,j,k)}}$.
2. If $\forall i \in \{1,\dots,d\}$, $\forall j \in \{1,\dots,n\}$, $\forall k \in \{1,\dots,m\}$: $f_{(i,j,k)}$ is differentiable at A, then f is differentiable at A.

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$: $B_n \to A$.

1. Let $i \in \{1,\dots,d\}$, $j \in \{1,\dots,n\}$, $k \in \{1,\dots,m\}$.

$$\frac{|f_{(i,j,k)}(B_n) - f_{(i,j,k)}(A) - \langle B_n - A, M_f(i,j,k)\rangle|}{\|B_n - A\|} \le \frac{\|f(B_n) - f(A) - (B_n - A)\odot M_f\|}{\|B_n - A\|} \to 0.$$

Hence $f_{(i,j,k)}$ is differentiable at A and $J_{f_{(i,j,k)}}(A) = M_f(i,j,k)$.

2. Set $M_f(i,j,k) = J_{f_{(i,j,k)}}(A)$. Then:

$$\frac{\|f(B_n) - f(A) - (B_n - A) \odot M_f\|}{\|B_n - A\|} = \left(\sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{m} \left(\frac{f_{(i,j,k)}(B_n) - f_{(i,j,k)}(A) - \langle B_n - A, J_{f_{(i,j,k)}}(A) \rangle}{\|B_n - A\|} \right.\right.$$
$$\to 0.$$

Hence f is differentiable at A. □

29 The Summation Operator and Hadamard Product on 3-Tensors

Definition 29.1 (Σ operator)**.** Let $A \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$. Then

$$\Sigma(A) \in \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}, \quad \Sigma(A) = \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{m} A_{(i,j,k)}.$$

Observation 29.1. Let $A, B \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$, $\lambda \in \mathbb{R}$. Then:

1. $\Sigma(A + B) = \Sigma(A) + \Sigma(B)$.
2. $\Sigma(\lambda A) = \lambda\Sigma(A)$.

Definition 29.2 (Hadamard-type product $*$)**.** Let $A \in \mathbb{R}^{d\times n\times m}$, $B \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$. Then

$$A * B \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}, \quad (A * B)_{(i,j,k)} = A_{(i,j,k)} B_{(i,j,k)}.$$

Definition 29.3 (Continuity of tensor-valued functions)**.** Let $X \subset \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : X \to \mathbb{R}^{d\times n\times m}$, $A \in X$. We say f is continuous at A if and only if

$$\forall \varepsilon > 0 : \exists \delta > 0 : B \in X, \|B - A\| < \delta \implies \|f(B) - f(A)\| < \varepsilon.$$

Equivalently: $B : \mathbb{N} \to X$, $B_n \to A \implies f(B_n) \to f(A)$.

Proposition 29.1. Let $X \subset \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : X \to \mathbb{R}^{d\times n\times m}$, $A \in X$. Then: f continuous at $A \iff \forall i \in \{1,\dots,d\}, \forall j \in \{1,\dots,n\}, \forall k \in \{1,\dots,m\}$: $f_{(i,j,k)}$ continuous at A.

Corollary 29.1. Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : \Omega \to \mathbb{R}^{d\times n\times m}$, $A \in \Omega$, f differentiable at A. Then f is continuous at A.

Proof. Since f is differentiable at A, we have that $\forall i \in \{1,\dots,d\}, \forall j \in \{1,\dots,n\}, \forall k \in \{1,\dots,m\}$: $f_{(i,j,k)}$ is differentiable at A, hence continuous at A. Therefore f is continuous at A. □

Proposition 29.2 (Submultiplicativity of $\odot$)**.** Let $A \in \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $B \in \left(\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}\right)^{d\times n\times m}$. Then: $\|A \odot B\| \le \|A\|\|B\|$.

Proof.

$$\begin{aligned}
\|A \odot B\| &= \left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}(A \odot B)_{(i,j,k)}^2\right)^{1/2} = \left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}\langle A, B_{(i,j,k)}\rangle^2\right)^{1/2} \\
&= \left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}\left(\sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} A_{(\hat{i},\hat{j},\hat{k})}B_{(i,j,k)}(\hat{i},\hat{j},\hat{k})\right)^2\right)^{1/2} \\
&\le \left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}\left(\sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} A_{(\hat{i},\hat{j},\hat{k})}^2\right)\left(\sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} B_{(i,j,k)}(\hat{i},\hat{j},\hat{k})^2\right)\right)^{1/2} \\
&= \left(\sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} A_{(\hat{i},\hat{j},\hat{k})}^2\right)^{1/2}\left(\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}\|B_{(i,j,k)}\|^2\right)^{1/2} = \|A\|\|B\|. \ \square
\end{aligned}$$

30 Boundedness Under Differentiability on $\mathbb{R}^{d\times n\times m}$

Proposition 30.1. Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : \Omega \to \mathbb{R}^{d\times n\times m}$, $A \in \Omega$, f differentiable at A, $B : \mathbb{N} \to \Omega\backslash\{A\}$: $B_n \to A$. Then the sequence $\alpha : \mathbb{N} \to \mathbb{R}$ with $\alpha_n = \|f(B_n) - f(A)\|/\|B_n - A\|$ is bounded.

Proof.

$$\begin{aligned}\alpha_n &\leq \frac{\|f(B_n) - f(A) - (B_n - A) \odot J_f(A)\|}{\|B_n - A\|} + \frac{\|(B_n - A) \odot J_f(A)\|}{\|B_n - A\|} \\ &= \delta_n + \frac{\|(B_n - A) \odot J_f(A)\|}{\|B_n - A\|} \leq \delta_n + \frac{\|B_n - A\|\|J_f(A)\|}{\|B_n - A\|} = \delta_n + \|J_f(A)\|.\end{aligned}$$

Since $\delta_n \to 0$, the sequence α is bounded. □

31 The Chain Rule on $\mathbb{R}^{d\times n\times m}$ (Scalar Case)

Theorem 31.1 (Chain Rule)**.** Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, G open in $\mathbb{R}^{d\times n\times m}$, $f : \Omega \to G$, $g : G \to \mathbb{R}$, $A \in \Omega$, f differentiable at A, g differentiable at $f(A)$. Then $(g \circ f) : \Omega \to \mathbb{R}$ is differentiable at A and:

$$\boxed{J_{g\circ f}(A) = \Sigma(J_g(f(A)) * J_f(A))\,.}$$

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$: $B_n \to A$.

$$\begin{aligned}&\frac{|(g \circ f)(B_n) - (g \circ f)(A) - \langle B_n - A, \Sigma(J_g(f(A)) * J_f(A))\rangle|}{\|B_n - A\|} \\ &\leq \frac{|g(f(B_n)) - g(f(A)) - \langle f(B_n) - f(A), J_g(f(A))\rangle|}{\|B_n - A\|} \\ &\quad + \frac{|\langle f(B_n) - f(A), J_g(f(A))\rangle - \langle B_n - A, \Sigma(J_g(f(A)) * J_f(A))\rangle|}{\|B_n - A\|} \\ &= a_n + b_n.\end{aligned}$$

For a_n: Define

$$a_n = \begin{cases} \dfrac{|g(f(B_n)) - g(f(A)) - \langle f(B_n) - f(A), J_g(f(A))\rangle|}{\|f(B_n) - f(A)\|} \cdot \dfrac{\|f(B_n) - f(A)\|}{\|B_n - A\|} & f(B_n) \neq f(A) \\ 0 & f(B_n) = f(A). \end{cases}$$

We have $a_n = \gamma_n \cdot \delta_n$ where $\gamma_n \to 0$ and δ_n is bounded. Hence $a_n \to 0$.

For b_n: We compute

$$\langle f(B_n) - f(A), J_g(f(A))\rangle = \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m}(f_{(i,j,k)}(B_n) - f_{(i,j,k)}(A))\frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A)).$$

Set $S = J_g(f(A)) * J_f(A)$, so $S_{(i,j,k)} = \dfrac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A)) \cdot J_{f_{(i,j,k)}}(A) \in \mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$.

Then $\Sigma(S)_{(\hat{i},\hat{j},\hat{k})} = \displaystyle\sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m} \frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A)) \cdot J_{f_{(i,j,k)}}(A)(\hat{i},\hat{j},\hat{k})$ and

$$\begin{aligned}\langle B_n - A, \Sigma(S)\rangle &= \sum_{\hat{i}}\sum_{\hat{j}}\sum_{\hat{k}} (B_n - A)_{(\hat{i},\hat{j},\hat{k})} \sum_{i}\sum_{j}\sum_{k} \frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A)) J_{f_{(i,j,k)}}(A)(\hat{i},\hat{j},\hat{k}) \\ &= \sum_{i}\sum_{j}\sum_{k} \frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A)) \langle B_n - A, J_{f_{(i,j,k)}}(A)\rangle.\end{aligned}$$

Therefore:

$$\begin{aligned}b_n &= \frac{1}{\|B_n - A\|}\left|\sum_{i}\sum_{j}\sum_{k}\left(f_{(i,j,k)}(B_n) - f_{(i,j,k)}(A) - \langle B_n - A, J_{f_{(i,j,k)}}(A)\rangle\right)\frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A))\right| \\ &\le \sum_{i}\sum_{j}\sum_{k}\left|\frac{\partial g}{\partial \hat{e}_{(i,j,k)}}(f(A))\right| \cdot \frac{|f_{(i,j,k)}(B_n) - f_{(i,j,k)}(A) - \langle B_n - A, J_{f_{(i,j,k)}}(A)\rangle|}{\|B_n - A\|} \to 0.\end{aligned}$$

□

32 The Scalar Chain Rule on $\mathbb{R}^{d\times n\times m}$

Proposition 32.1. Let Ω be open in $\mathbb{R}^{d\times n\times m}$, $f : \Omega \to \mathbb{R}$, $A \in \Omega$, f differentiable at A, $g : \mathbb{R} \to \mathbb{R}$ differentiable. Then the composition $(g \circ f) : \Omega \to \mathbb{R}$ is differentiable at A and:

$$J_{g\circ f}(A) = g'(f(A))\, J_f(A).$$

Proof. Let $B : \mathbb{N} \to \Omega \setminus \{A\}$: $B_n \to A$. Then:

$$\lim_{n\to\infty} \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} = 0.$$

We have:

$$\begin{aligned}&\frac{|(g\circ f)(B_n) - (g\circ f)(A) - \langle B_n - A, g'(f(A))J_f(A)\rangle|}{\|B_n - A\|} \\ &\le \frac{|g(f(B_n)) - g(f(A)) - (f(B_n) - f(A))g'(f(A))|}{\|B_n - A\|} \\ &\quad + \frac{|(f(B_n) - f(A))g'(f(A)) - \langle B_n - A, g'(f(A))J_f(A)\rangle|}{\|B_n - A\|} = a_n + b_n.\end{aligned}$$

For a_n: Define

$$a_n = \begin{cases} \left| \dfrac{g(f(B_n)) - g(f(A))}{f(B_n) - f(A)} - g'(f(A)) \right| \cdot \dfrac{|f(B_n) - f(A)|}{\|B_n - A\|} & f(B_n) \neq f(A) \\ 0 & f(B_n) = f(A). \end{cases}$$

Claim: $\delta_n = |f(B_n) - f(A)|/\|B_n - A\|$ is bounded.
Indeed:

$$\begin{aligned} \delta_n &\leq \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} + \frac{|\langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \\ &\leq \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} + \|J_f(A)\|. \end{aligned}$$

Since the first term tends to 0, δ_n is bounded. Since $\gamma_n \to 0$, we get $a_n \to 0$.
For b_n:

$$b_n = |g'(f(A))| \cdot \frac{|f(B_n) - f(A) - \langle B_n - A, J_f(A)\rangle|}{\|B_n - A\|} \to 0. \qquad \square$$

33 Element-wise Composition on $\mathbb{R}^{d\times n\times m}$

Corollary 33.1. Let Ω be open in $\mathbb{R}^{\hat{d}\times\hat{n}\times\hat{m}}$, $f : \Omega \to \mathbb{R}^{d\times n\times m}$, $A \in \Omega$, f differentiable at A, $g : \mathbb{R} \to \mathbb{R}$ differentiable. Set $g\langle f\rangle : \Omega \to \mathbb{R}^{d\times n\times m}$ with $g\langle f\rangle_{(i,j,k)} = g \circ f_{(i,j,k)}$. Then $g\langle f\rangle$ is differentiable at A and:

$$J_{g\langle f\rangle}(A) = g'[f(A)] * J_f(A)$$

where $g'[f(A)] \in \mathbb{R}^{d\times n\times m}$ with $g'[f(A)]_{(i,j,k)} = g'(f_{(i,j,k)}(A))$.

Proof. Since f is differentiable at A, we have that $\forall i \in \{1,\dots,d\}$, $\forall j \in \{1,\dots,n\}$, $\forall k \in \{1,\dots,m\}$: $f_{(i,j,k)}$ is differentiable at A. Hence $\forall i,j,k$: $g\circ f_{(i,j,k)} : \Omega \to \mathbb{R}$ is differentiable at A and $J_{g\circ f_{(i,j,k)}} = g'(f_{(i,j,k)}(A))J_{f_{(i,j,k)}}(A)$. Therefore $g\langle f\rangle$ is differentiable at A and:

$$J_{g\langle f\rangle}(A) = g'[f(A)] * J_f(A). \qquad \square$$

34 Slicing, Submatrix Extraction, and Convolution Operations

Definition 34.1 (Slicing)**.** Let $D \in \mathbb{R}^{d\times n\times m}$. For $i \in \{1,\dots,d\}$, we define the *slice* $D^{(i)} \in \mathbb{R}^{n\times m}$ by $D^{(i)}_{(j,k)} = D_{(i,j,k)}$.

Definition 34.2 (Submatrix extraction)**.** Let $A \in \mathbb{R}^{n\times m}$, $i, j, \hat{n}, \hat{m} \in \mathbb{N}^*$: $i + \hat{n} \le n$, $j + \hat{m} \le m$. Then $A\langle i, j, \hat{n}, \hat{m}\rangle \in \mathbb{R}^{\hat{n}\times\hat{m}}$ with

$$A\langle i, j, \hat{n}, \hat{m}\rangle(\tilde{i}, \tilde{j}) = A(i + \tilde{i}, j + \tilde{j}).$$

Definition 34.3 (2D Convolution)**.** Let $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{\hat{n}\times\hat{m}}$; $\hat{n} \le n$, $\hat{m} \le m$. Then $A\textcircled{C}B \in \mathbb{R}^{(n-\hat{n}+1)\times(m-\hat{m}+1)}$ with

$$(A\textcircled{C}B)_{(i,j)} = \langle A\langle i, j, \hat{n}, \hat{m}\rangle, B\rangle.$$

Definition 34.4 (Tensor Convolution)**.** Let $A \in \mathbb{R}^{d\times n\times m}$, $B \in \mathbb{R}^{1\times\hat{n}\times\hat{m}}$; $\hat{n} \le n$, $\hat{m} \le m$. Then $A\textcircled{C}B \in \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$ with

$$(A\textcircled{C}B)_{(i,j,k)} = \langle A^{(i)}\langle j, k, \hat{n}, \hat{m}\rangle, B\rangle.$$

That is, $(A\textcircled{C}B)^{(i)} = A^{(i)}\textcircled{C}B$.

35 Application: Convolution with Fixed Filter

Definition 35.1. Let $\omega \in \mathbb{R}^{\hat{n}\times\hat{m}}$. Define $M(\omega) \in \mathbb{R}^{\hat{n}\times\hat{m}}$ by $M(\omega)_{(i,j)} = \omega_{(\hat{n}-i+1,\hat{m}-j+1)}$.

Proposition 35.1. Let $\omega \in \mathbb{R}^{\hat{n}\times\hat{m}}$. Define $f : \mathbb{R}^{d\times n\times m} \to \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$ by $f(x) = x\textcircled{C}\omega$.

Then $J_f(D) \in \left(\mathbb{R}^{d\times n\times m}\right)^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$.

For $i \in \{1,\dots,d\}$, $j \in \{1,\dots,n-\hat{n}+1\}$, $k \in \{1,\dots,m-\hat{m}+1\}$:

$$f_{(i,j,k)}(x) = (x\textcircled{C}\omega)_{(i,j,k)} = \langle x^{(i)}\langle j, k, \hat{n}, \hat{m}\rangle, \omega\rangle$$

and the partial derivative is:

$$\frac{\partial f_{(i,j,k)}}{\partial \hat{e}_{(\hat{i},\hat{j},\hat{k})}}(D) = \begin{cases} 0 & \hat{i} \neq i \\ \omega_{(\hat{j}-j+1,\hat{k}-k+1)} & \hat{i} = i,\ \hat{j} \in \{j,\dots,\hat{n}+j-1\},\ \hat{k} \in \{k,\dots,\hat{m}+k-1\} \\ 0 & \text{otherwise.} \end{cases}$$

Hence:

$$J_{f_{(i,j,k)}}(D) = e^{d\times(n+\hat{n}-1)\times(m+\hat{m}-1)}_{(i,\hat{n}+j-1,\hat{m}+k-1)} \textcircled{C} M(\omega).$$

Corollary 35.1. Let $\Lambda \in \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$.

$$\Sigma(\Lambda * J_f(D)) = \hat{\Lambda} \textcircled{C} M(\omega)$$

where $\hat{\Lambda} \in \mathbb{R}^{d\times(n+\hat{n}-1)\times(m+\hat{m}-1)}$ is the zero-padded version of Λ:

$$\hat{\Lambda}^{(i)} = \sum_{j=1}^{n-\hat{n}+1} \sum_{k=1}^{m-\hat{m}+1} \Lambda_{(i,j,k)}\, e^{(n+\hat{n}-1)\times(m+\hat{m}-1)}_{(\hat{n}+j-1,\hat{m}+k-1)}.$$

36 Application: Convolution with Fixed Input

Proposition 36.1. Let $D \in \mathbb{R}^{d\times n\times m}$. Define $f : \mathbb{R}^{\hat{n}\times\hat{m}} \to \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$ by $f(x) = D \textcircled{C} x$.

Then $J_f(\omega) \in \left(\mathbb{R}^{\hat{n}\times\hat{m}}\right)^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$.

For $i \in \{1,\ldots,d\}$, $j \in \{1,\ldots,n-\hat{n}+1\}$, $k \in \{1,\ldots,m-\hat{m}+1\}$:

$$f_{(i,j,k)}(x) = (D \textcircled{C} x)_{(i,j,k)} = \langle D^{(i)} \langle j,k,\hat{n},\hat{m}\rangle, x\rangle$$

and:

$$\frac{\partial f_{(i,j,k)}}{\partial \hat{e}_{(\hat{j},\hat{k})}}(\omega) = D^{(i)}_{(j+\hat{j}-1,k+\hat{k}-1)}.$$

Hence:

$$J_{f_{(i,j,k)}}(\omega) = D^{(i)} \textcircled{C} e^{(n-\hat{n}+1)\times(m-\hat{m}+1)}_{(j,k)}.$$

Corollary 36.1. Let $\Lambda \in \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$. Then:

$$\Sigma(\Lambda * J_f(\omega)) = \sum_{i=1}^{d} D^{(i)} \textcircled{C} \Lambda^{(i)}.$$

37 Bias Addition

Definition 37.1. Let $D \in \mathbb{R}^{d\times n\times m}$, $b \in \mathbb{R}^{n\times m}$. Then $D \oplus b \in \mathbb{R}^{d\times n\times m}$: $(D \oplus b)^{(i)} = D^{(i)} + b$.

Proposition 37.1. Let $D \in \mathbb{R}^{d\times n\times m}$. Define $f : \mathbb{R}^{n\times m} \to \mathbb{R}^{d\times n\times m}$ by $f(x) = D \oplus x$.

Then $J_f(b) \in (\mathbb{R}^{n\times m})^{d\times n\times m}$. For all $i \in \{1,\dots,d\}$, $j \in \{1,\dots,n\}$, $k \in \{1,\dots,m\}$:

$$f_{(i,j,k)}(x) = D^{(i)}_{(j,k)} + x_{(j,k)}, \quad \frac{\partial f_{(i,j,k)}}{\partial \hat{e}_{(j,k)}}(b) = 1.$$

Hence each component of $J_f(b)$ is the identity-like matrix.

Corollary 37.1. Let $\Lambda \in \mathbb{R}^{d\times n\times m}$. Then:

$$\Sigma(\Lambda * J_f(b)) = \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=1}^{m} \Lambda_{(i,j,k)} \cdot \mathbf{1}_{(j,k)} = \Sigma(\Lambda) \cdot \mathbf{1}^{n\times m}.$$

38 Application to Convolutional Neural Networks

We now apply the 3-tensor differentiability theory to convolutional neural networks [LBD+89, GBC16].

Let ω, D, b be known. Let $J_E(Y) \in \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}$ be known. Set $Y = \sigma((D \text{Ⓒ} \omega) \oplus b)$.

38.1 Gradient with respect to D

Define:

$$\begin{aligned}
&f_1 : \mathbb{R}^{d\times n\times m} \to \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}, && f_1(x) = (x \text{Ⓒ} \omega) \oplus b,\\
&f_2 : \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)} \to \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)}, && f_2 = \sigma\langle f_1\rangle,\\
&f_D : \mathbb{R}^{d\times n\times m} \to \mathbb{R}, && f_D = E \circ f_2.
\end{aligned}$$

Then:

$$J_{f_D}(D) = \Sigma(J_E(f_2(D)) * J_{f_2}(D)) = \Sigma(J_E(Y) * J_{f_2}(D))\,.$$

Since $J_{f_2}(D) = [Y(1-Y)] * J_{f_1}(D)$, we get:

$$\begin{aligned} J_{f_D}(D) &= \Sigma(J_E(Y) * ([Y(1-Y)] * J_{f_1}(D))) \\ &= \Sigma([J_E(Y) \cdot Y(1-Y)] * J_{f_1}(D)) \end{aligned}$$

$$\boxed{J_{f_D}(D) = \hat{M} \textcircled{C} M(\omega), \quad \text{where } M = J_E(Y) \cdot Y(1-Y).}$$

38.2 Gradient with respect to ω

Define:

$$\begin{aligned} &g_1 : \mathbb{R}^{\hat{n} \times \hat{m}} \to \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)}, && g_1(x) = [D \textcircled{C} x] \oplus b, \\ &g_2 : \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)} \to \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)}, && g_2 = \sigma\langle g_1 \rangle, \\ &g_\omega : \mathbb{R}^{\hat{n} \times \hat{m}} \to \mathbb{R}, && g_\omega = E \circ g_2. \end{aligned}$$

Then:

$$J_{g_\omega}(\omega) = \Sigma(J_E(g_2(\omega)) * J_{g_2}(\omega)) = \Sigma(J_E(Y) * J_{g_2}(\omega)) \,.$$

Since $J_{g_2}(\omega) = [Y(1-Y)] * J_{g_1}(\omega)$:

$$J_{g_\omega}(\omega) = \Sigma([J_E(Y) \cdot Y(1-Y)] * J_{g_1}(\omega))$$

$$\boxed{J_{g_\omega}(\omega) = \sum_{i=1}^{d} D^{(i)} \textcircled{C} M^{(i)}, \quad \text{where } M = J_E(Y) \cdot Y(1-Y).}$$

38.3 Gradient with respect to b

Define:

$$\begin{aligned} &h_1 : \mathbb{R}^{(n-\hat{n}+1) \times (m-\hat{m}+1)} \to \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)}, && h_1(x) = [D \textcircled{C} \omega] \oplus x, \\ &h_2 : \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)} \to \mathbb{R}^{d \times (n-\hat{n}+1) \times (m-\hat{m}+1)}, && h_2 = \sigma\langle h_1 \rangle, \\ &h_b : \mathbb{R}^{(n-\hat{n}+1) \times (m-\hat{m}+1)} \to \mathbb{R}, && h_b = E \circ h_2. \end{aligned}$$

Then:

$$\begin{aligned} J_{h_b}(b) &= \Sigma(J_E(h_2(b)) * J_{h_2}(b)) = \Sigma(J_E(Y) * [Y(1-Y)] * J_{h_1}(b)) \\ &= \Sigma([J_E(Y) \cdot Y(1-Y)] * J_{h_1}(b)) \end{aligned}$$

$$\boxed{J_{h_b}(b) = \Sigma(M) \cdot \mathbf{1}^{(n-\hat{n}+1) \times (m-\hat{m}+1)}, \quad \text{where } M = J_E(Y) \cdot Y(1-Y).}$$

References

[Cau21] Augustin-Louis Cauchy. *Cours d'analyse de l'École Royale Polytechnique*. Imprimerie Royale, 1821. The Cauchy–Schwarz inequality.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[HJ12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[LBD+89] Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[Sch85] Hermann Amandus Schwarz. Über ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung. *Acta Societatis Scientiarum Fennicae*, 15:315–362, 1885.

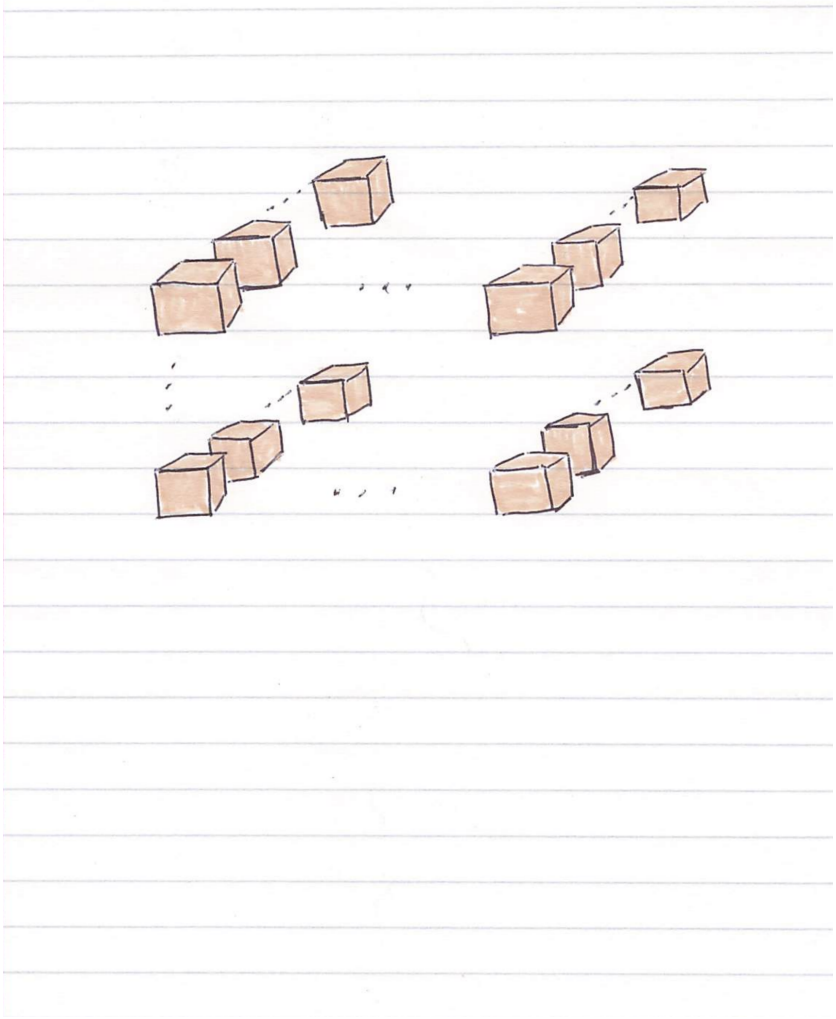

Figure 1: Tensor visualization showing matrices arranged as $n \times m$ grid of blocks

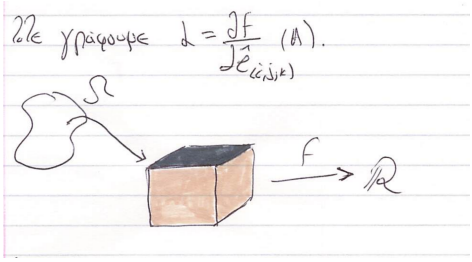

Figure 2: Visualization of a 3-tensor $A \in \mathbb{R}^{d \times n \times m}$: a "box" mapping into $\mathbb{R}$ via f.

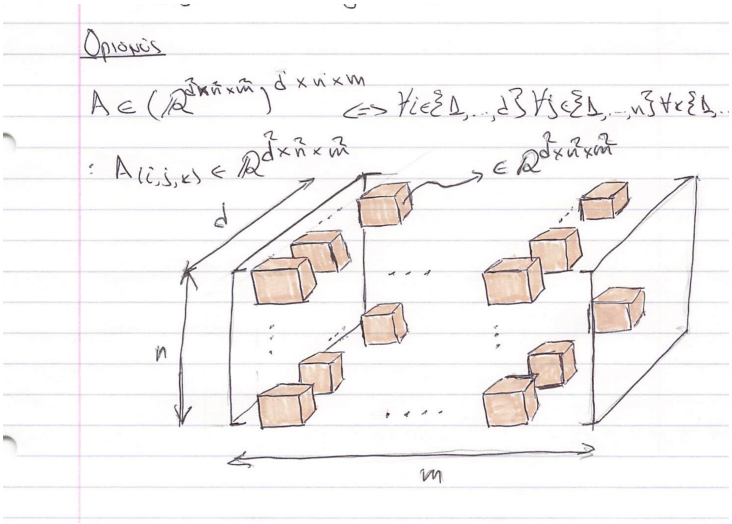

Figure 3: Visualization of a tensor $A \in \left(\mathbb{R}^{\hat{d} \times \hat{n} \times \hat{m}}\right)^{d \times n \times m}$: a $d \times n \times m$ array whose entries are themselves $\hat{d} \times \hat{n} \times \hat{m}$ tensors.

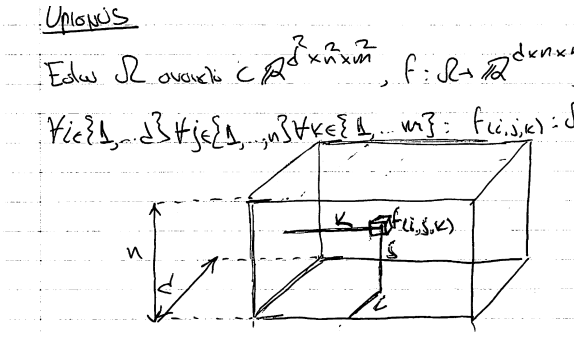

Figure 4: A tensor-valued function $f : \Omega \to \mathbb{R}^{d \times n \times m}$, with component functions $f_{(i,j,k)} : \Omega \to \mathbb{R}$.

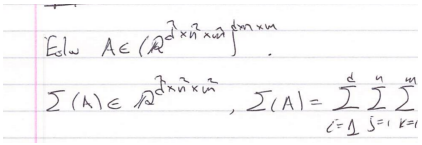

Figure 5: The summation operator Σ: summing all component tensors $A_{(i,j,k)}$.

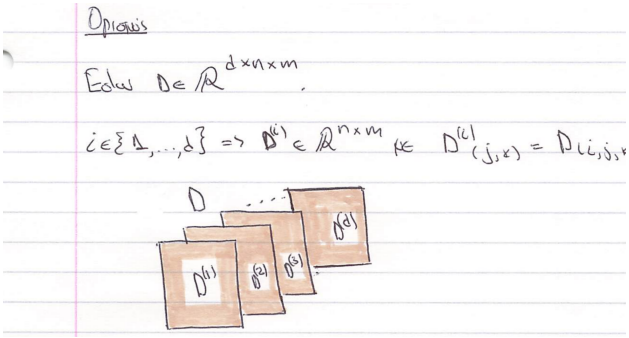

Figure 6: Slicing a 3-tensor $D \in \mathbb{R}^{d \times n \times m}$ into d matrices $D^{(1)}, D^{(2)}, \ldots, D^{(d)}$, each in $\mathbb{R}^{n \times m}$.

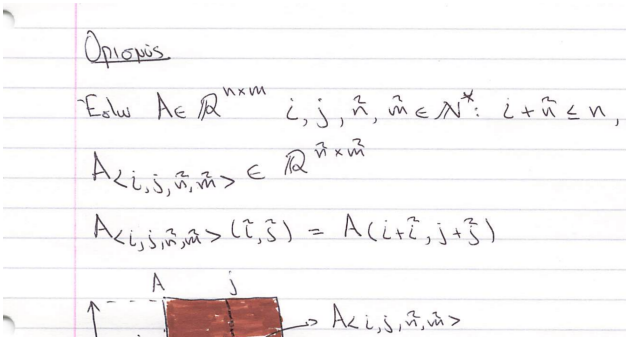
Ορισμός
Έστω $A \in \mathbb{R}^{n \times m}$ $i, j, \hat{n}, \hat{m} \in \mathbb{N}^*$: $i + \hat{n} \leq n$,
$A\langle i, j, \hat{n}, \hat{m} \rangle \in \mathbb{R}^{\hat{n} \times \hat{m}}$
$A\langle i, j, \hat{n}, \hat{m} \rangle (\hat{i}, \hat{j}) = A(i + \hat{i}, j + \hat{j})$
A
j
$A\langle i, j, \hat{n}, \hat{m} \rangle$

Figure 7: Submatrix extraction: the submatrix $A\langle i, j, \hat{n}, \hat{m} \rangle$ is a window of size $\hat{n} \times \hat{m}$ starting at position (i, j) in the matrix A.

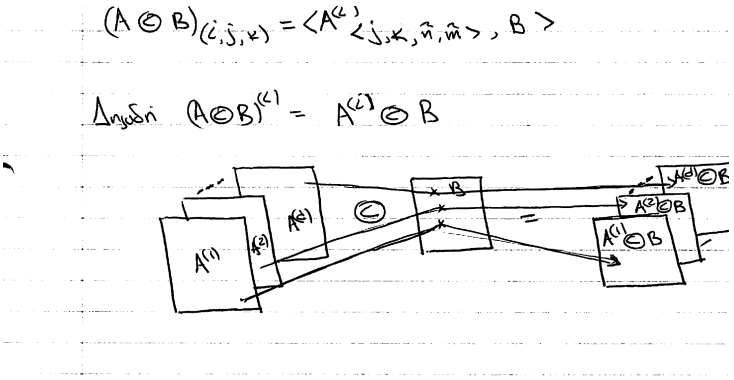

Figure 8: Tensor convolution: each slice $A^{(i)}$ is convolved with the filter B, producing a slice of the output tensor.

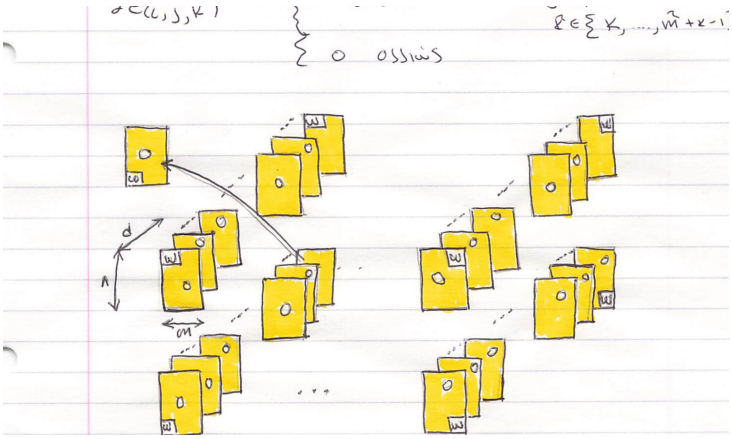

Figure 9: Visualization of the Jacobian tensor $J_f(D)$: each component is a convolution of a standard basis tensor with the rotated filter $M(\omega)$.

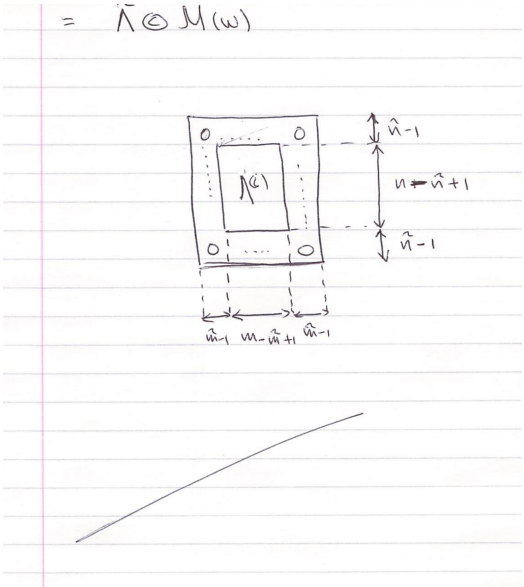

Figure 10: Zero-padding of $\Lambda^{(i)}$: the original $(n-\hat{n}+1)\times(m-\hat{m}+1)$ matrix is embedded in a larger $(n+\hat{n}-1)\times(m+\hat{m}-1)$ matrix with $(\hat{n}-1)$ zeros on each side.

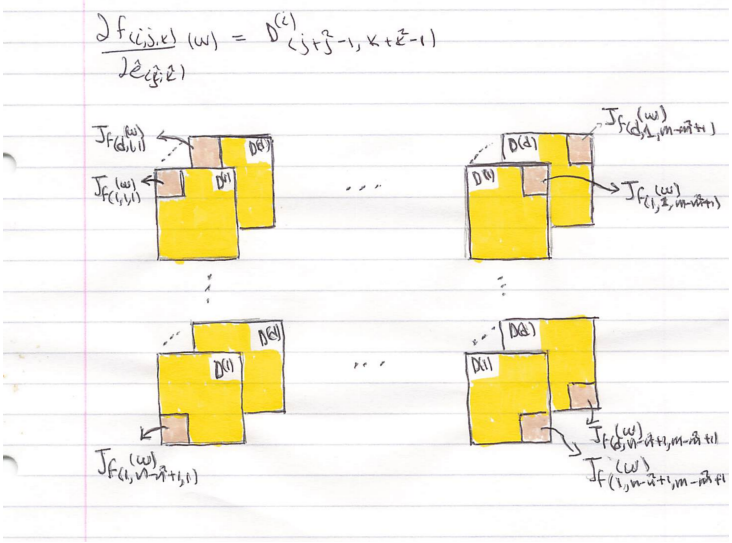

Figure 11: Visualization of the Jacobian components $J_{f_{(i,j,k)}}(\omega)$: each is a submatrix of $D^{(i)}$.

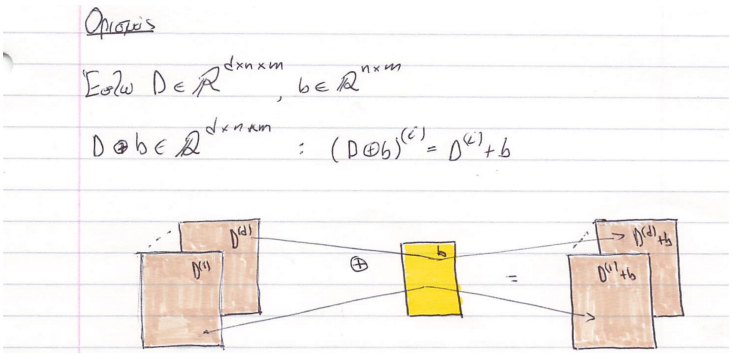

Figure 12: Bias addition: each slice $D^{(i)}$ has the same bias matrix b added to it.

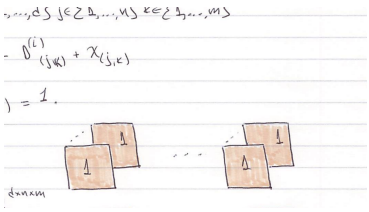

Figure 13: The Jacobian of bias addition: each component is a matrix of ones at the corresponding position.

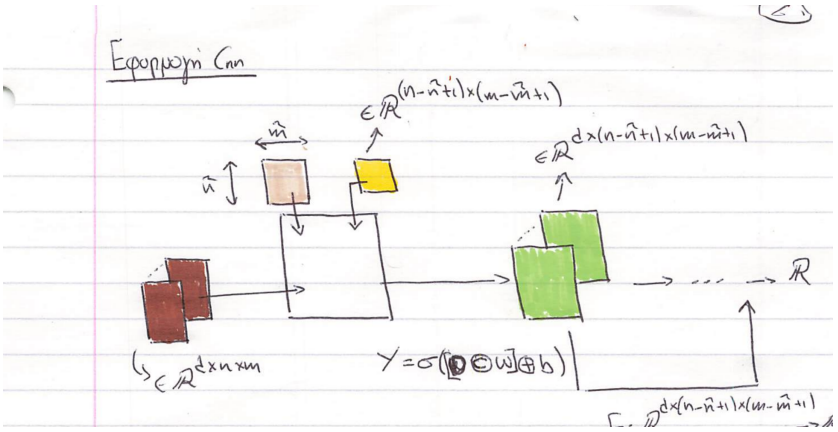

Figure 14: A convolutional neural network layer: the input $D \in \mathbb{R}^{d\times n\times m}$ is convolved with filter $\omega \in \mathbb{R}^{\hat{n}\times\hat{m}}$, bias b is added, and an activation function σ is applied element-wise. The output $Y = \sigma((D \textcircled{C} \omega) \oplus b)$ feeds into the next layer and eventually into a loss function $E : \mathbb{R}^{d\times(n-\hat{n}+1)\times(m-\hat{m}+1)} \to \mathbb{R}$.